DTIC FILE COPY

(2)

AD-A194 448

20000920167

DTIC
ELECTE
JUN 09 1988
S D
H

AIR COMMAND AND STAFF COLLEGE

STUDENT REPORT

THE OPERATIONAL TESTING EFFECTIVENESS EVALUATION METHOD

MAJOR WILHELM F. PERCIVAL 88-2090

"insights into tomorrow"

Reproduced From
Best Available Copy

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

88 6 6 070

DISCLAIMER

The views and conclusions expressed in this document are those of the author. They are not intended and should not be thought to represent official ideas, attitudes, or policies of any agency of the United States Government. The author has not had special access to official information or ideas and has employed only open-source material available to any writer on this subject.

This document is the property of the United States Government. It is available for distribution to the general public. A loan copy of the document may be obtained from the Air University Interlibrary Loan Service (AUL/LDEX, Maxwell AFB, Alabama, 36112-5564) or the Defense Technical Information Center. Request must include the author's name and complete title of the study.

This document may be reproduced for use in other research reports or educational pursuits contingent upon the following stipulations:

- Reproduction rights do not extend to any copyrighted material that may be contained in the research report.

- All reproduced copies must contain the following credit line: "Reprinted by permission of the Air Command and Staff College."

- All reproduced copies must contain the name(s) of the report's author(s).

- If format modification is necessary to better serve the user's needs, adjustments may be made to this report--this authorization does not extend to copyrighted information or material. The following statement must accompany the modified document: "Adapted from Air Command and Staff College Research Report (number) entitled (title) by (author)."

- This notice must be included with any reproduced or adapted portions of this document.

REPORT NUMBER 88-2090

TITLE THE OPERATIONAL TESTING EFFECTIVENESS EVALUATION METHOD

AUTHOR(S) MAJOR WILHELM F. PERCIVAL, USAF

FACULTY ADVISOR MAJOR LARRY J. PULCHER, ACSC/EPT

SPONSOR LT COL ROBERT F. BEHLER
31st TEST AND EVALUATION SQUADRON (SAC)
31 TES/CC
EDWARDS AFB, CA 93523

Submitted to the faculty in partial fulfillment of requirements for graduation.

AIR COMMAND AND STAFF COLLEGE
AIR UNIVERSITY
MAXWELL AFB, AL 36112-5542

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

ADA194448

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-01

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT STATEMENT "A" Approved for public release; Distribution is unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) 88-2090	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION ACSC/EDC 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Maxwell AFB AL 36112-5542	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING / SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO / WORK UNIT ACCESSION

11. TITLE (Include Security Classification)

THE OPERATIONAL TESTING EFFECTIVENESS EVALUATION METHOD

12. PERSONAL AUTHOR(S)

Percival, Wilhelm F., Major, USAF

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
	FROM _____ TO _____	1988 April	40

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Despite the critical role played by Operational Test and Evaluation (OT&E) in ensuring the effectiveness and suitability of new weapons, no procedure exists for objectively measuring testing effectiveness. This lack of feedback is particularly unfortunate in the case of Initial Operational Test and Evaluation (IOT&E), where an incorrect negative assessment can kill a new weapons system. This report examines the need for an operational testing effectiveness evaluation procedure, first discussing recent weapons testing controversies, and then reviewing OT&E history. Present-day OT&E missions and challenges are reviewed prior to introduction of the proposed measurement technique--the Operational Testing Effectiveness Evaluation Method (OTEEM). The new method, designed to measure the adequacy and accuracy of IOT&E, is then applied to an actual weapons system IOT&E program. After discussion of problems and concerns, implementation of OTEEM is recommended.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL ACSC/EDC Maxwell AFB AL 36112-5542	22b. TELEPHONE (Include Area Code) (205) 293-2867 / 22c. OFFICE SYMBOL

DD Form 1473, JUN 86 *Previous editions are obsolete.*

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

PREFACE

The interpretation and significance of test results is a common subject of contention in test programs. For example, Initial Operational Test and Evaluation (IOT&E) personnel often debate how particular test results should affect the operational effectiveness or suitability assessment. In this debate, the battle lines are often drawn by organization--the program office on one side, the test community on the other. Sometimes, using the same test data, engineers from the two organizations reach radically different conclusions. Which group would be proved right when the weapon was used in the field? The question of how test assessments stack up against a weapon's later operational performance has far-reaching implications.

Today, IOT&E assessments play a critical role in acquisition decision making. Given their importance, how accurate are these assessments? For example, did testers accurately predict maintainability (fuel leak) problems with the B-1B or effectiveness deficiencies with the Division Air Defense (DIVAD) gun? For operational weapons systems, it might be possible to check assessment accuracy by comparing the IOT&E assessments against actual performance data gathered in the field. Such a comparison could reveal whether IOT&E assessments were ultimately right or wrong--feedback that should have all sorts of valuable applications. For example, diverse IOT&E programs could be rated for assessment accuracy and compared, testing methods improved, and critics silenced. Given that checking IOT&E assessments against operational data seemed to be common sense, was somebody was already doing it?

After checking with the Air Force Operational Test and Evaluation Center (AFOTEC) and the OSD office for OT&E, it became clear that IOT&E assessments are never compared to a weapon's later operational performance. (21:--; 22:--) Nobody ever looks back at the IOT&E results to check accuracy. This report is an attempt to fill this void with a feedback tool for IOT&E called the Operational Testing Effectiveness Evaluation Method (OTEEM). This is virgin ground and the work in this report is really only a starting point. Changes will undoubtedly be made, but the idea is to get the ball rolling toward eventual implementation of this potentially valuable idea.

Special thanks go to my advisor, Maj Larry Pulcher, who helped me achieve some degree of coherence in this paper.

DTIC COPY INSPECTED

Accession For	
NTIS GRA&I	☑
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

ABOUT THE AUTHOR

Major Percival has a Strategic Air Command (SAC) background. After completing Combat Crew Training School in 1977, he was assigned to the 416th Bomb Wing, Griffiss AFB, NY, where he eventually became an instructor/evaluator pilot in the B-52G. He participated actively in the first operational deployment of the Air Launched Cruise Missile and B-52 Offensive Avionics System (OAS). He was then assigned to the 31st Test and Evaluation Squadron (SAC) at Edwards AFB, CA, where he worked on several projects including the OAS Block II software test, the Common Strategic Rotary Launcher, and the Advanced Cruise Missile (ACM). As a member of the Strategic Systems Combined Test Force, he was the senior B-52 flight examiner, Deputy for Operations, and Deputy Test Director for the ACM Initial Operational Test and Evaluation test force. Currently, he is a student at the Air Command and Staff College, Maxwell AFB, AL. After graduation in 1988, he will return to the 31st Test and Evaluation Squadron to work on another test program.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

TABLES

EXECUTIVE SUMMARY

Part of our College mission is distribution of the students' problem solving products to DoD sponsors and other interested agencies to enhance insight into contemporary, defense related issues. While the College has accepted this product as meeting academic requirements for graduation, the views and opinions expressed or implied are solely those of the author and should not be construed as carrying official sanction.

"insights into tomorrow"

REPORT NUMBER 88-2090

AUTHOR(S) MAJOR WILHELM F. PERCIVAL

TITLE THE OPERATIONAL TESTING EFFECTIVENESS EVALUATION METHOD

I. Problem: Operational Test and Evaluation (OT&E) plays a vital role in weapons system acquisition. Decision makers rely on Initial Operational Test and Evaluation (IOT&E) effectiveness and suitability assessments when making acquisition decisions. The problem is that there is currently no attempt to check the accuracy or adequacy of these OT&E assessments--in short, no way to evaluate OT&E effectiveness.

II. Objectives: The objective of this report is to support the need for feedback in operational testing and introduce a technique designed to measure IOT&E effectiveness. This proposed technique is called the Operational Testing Effectiveness Evaluation Method (OTEEM). The report will show that OTEEM should be implemented.

III. Discussion: The first two chapters of the report support the need for OTEEM. Chapter One shows how critics have disputed the adequacy and accuracy of testing in such systems as the B-1B, the Division Air Defense Gun, and the Advanced Medium-Range Air-to-Air Missile. These disputes highlight the need for an objective evaluation of operational testing effectiveness. The second chapter reviews the history of OT&E, discussing the possibility that the record of frequent convulsive organizational change is related to the lack of adequate feedback on OT&E effectiveness. Furthermore, without objective feedback, today's managers may continue the historical pattern of ineffective

change. The third chapter lays the foundation for OTEEM by reviewing the present mission of OT&E and the challenges OT&E personnel must face. The chapter is intended for those unfamiliar with OT&E. Chapter Four is the crux of the report, as it introduces and applies OTEEM to the Air Launched Cruise Missile (ALCM). OTEEM relies on a comparison of the IOT&E assessments made in the IOT&E final report and the results of field testing summarized in the FOT&E Phase One final report. The method compares the areas of operational effectiveness, suitability, critical issue assessment, and deficiency reporting. Survivability is mentioned as an OTEEM assessment area, but is not included in the example due to classification. The ALCM example serves to illustrate the OTEEM technique and suggest improvements or problems. Chapter Five discusses several findings, including problems and concerns raised by the application example.

IV. Findings: The application exercise shows that OTEEM is capable of uncovering problems in IOT&E. The in-depth OTEEM analysis of a test program provides valuable insights for the OT&E manager. In addition to scrutinizing individual programs, the manager would be able to summarize and compare numerous test programs to assess broad trends in operational testing. Other benefits would include the fine-tuning of effectiveness and suitability forecasting techniques and the identification of common pitfalls for OT&E managers to avoid. Finally, several minor improvements in final report format or approach would facilitate OTEEM application. Overall, OTEEM seems to offer significant benefits--including increased confidence in OT&E assessments--for minimal cost.

V. Recommendation: The Air Force Operational Test and Evaluation Center (AFOTEC) should begin a trial OTEEM application program. After this trial period, a finalized form of OTEEM should be implemented.

CHAPTER ONE

THE NEED

> So in war, through the influence of innumerable trifling circumstances, which on paper cannot properly be taken into consideration, everything depresses us and we come far short of our mark. (19:45)
>
> - Clausewitz

> Our weapons tests now use so much computer modeling and simulation that no one knows whether some new arms really work. (4:50)
>
> - Discover Magazine

The debate rages in the press, in technical journals, in the halls of the Pentagon and Congress, and in the crew lounges of operational squadrons. Will the new high-tech weapons work in combat, or even in peacetime? Before it buys these weapons, the Air Force tests them to answer that question. Therefore, poor weapon effectiveness, if it exists, can be intimately linked to poor testing effectiveness. Currently, the Air Force has no way of objectively and routinely judging the effectiveness of weapons system testing. This report is about a method designed to provide objective feedback on the effectiveness of Initial Operational Test and Evaluation (IOT&E). The first two chapters establish the need for this method. Chapters Three and Four develop the method and demonstrate its use. The final two chapters examine the issues raised by earlier chapters, summarize the report, and recommend action.

As the title states, this chapter is about the need. In an Air Force where every conceivable performance dimension is measured, it seems odd to argue for more feedback. However, IOT&E is an area where objective feedback is critically needed. To appreciate why the Air Force needs to evaluate the effectiveness of operational testing, it helps to review the official purpose of test and evaluation and then contrast its utopian wording with some short examples of real-world controversy.

Over the years, the Air Force established test and evaluation procedures to find out if weapons work. According to Air Force Regulation 80-14, the purpose of all test and evaluation is: "to identify, assess, and reduce the acquisition risks; to evaluate operational effectiveness and operational suitability; to identify any deficiencies in the system; and to ensure that only operationally effective and suitable, supportable systems are delivered to the operating forces." (17:2) In other words, testing determines if weapons work as advertised and forecasts their effectiveness on the battlefield. Furthermore, testing ensures that only effective and suitable weapons make it to the ramp. Sounds easy, but as the following examples imply, the testing job is much more difficult than it appears.

The nation's newest strategic bomber, the B-1B, is flying through a storm of controversy surrounding its operational capabilities. In recent months, the aircraft has received negative press on problems ranging from fuel leaks to faulty defensive avionics. (10:--) B-1 supporters contend the aircraft is just experiencing "routine" difficulties; nothing to be alarmed about. (9:--) However, with articles like "The B-1 Bomber: A Flying Lemon?" spreading alarm seems to be the media's goal. (9:--) The Air Force Chief of Staff, responding to the feeding-frenzy atmosphere generated by B-1 critics, has complained about ". . . hypercritical reports in the media, even in such level-headed places as Texas." (3:--) Meanwhile, testifying to congressional subcommittees, "Gen. Lawrence Welch admitted that the Air Force failed to adequately test major B-1B subsystems before they were integrated into the aircraft." (8:264) Questions of adequacy and objectivity have also dogged other Department of Defense (DoD) test programs.

One such program, the Division Air Defense Gun (DIVAD), is significant because the DoD directives governing the Army's DIVAD testing also govern Air Force testing. DIVAD, or Sgt York, was ". . . the first major weapons system to be scrubbed in eight years and the first in decades to be canceled so far into production." (6:--) When the system was canceled, a significant amount of testing had already been performed. According to the DoD directives current at the time entry into Full Scale Development (FSD) required "adequate" developmental and operational testing to identify risks, "feasible solutions," and "estimate the potential operational effectiveness. . . ." (16:13-14) Unfortunately, some of this testing was apparently rigged in DIVAD's favor. (4:56) However, the subsequent operational testing required for Low Rate Initial Production (LRIP) finally and conclusively showed the weapon was a flop. (7:44) Therefore, after experiencing initial difficulties, IOT&E successfully revealed DIVAD's problems.

The senior executive charged with the operational evaluation of new weapons, OSD's Director of Operational Test and Evaluation (DOTE), proved the value of independent OT&E by blowing the whistle on DIVAD. But even then, he had to respond to accusations of soft-pedalling DIVAD problems. (7:46) Apparently, some politicians doubted DOTE's objectivity. Writing in 1986, Senator Gary Hart said: "It [DOTE] is playing the same 'go along to get along, keep everybody happy by keeping the money flowing' game that has too often undermined past operational testing and effective weapons." (7:42) In the end, Sgt York cost 1.8 billion dollars and, according to some, another black eye for weapons acquisition and testing. (4:56) Certain critics think the Advanced Medium-Range Air-to-Air Missile (AMRAAM) could be another DIVAD. (7:46)

The AMRAAM was certified for LRIP on 28 Feb 86, despite DOTE memoranda warning there was a "low probability of adequate test results" being obtained prior to certification. (7:46) Simultaneously, a General Accounting Office report critical of the missile added fuel to the fire of AMRAAM critics on Capitol Hill. (7:46) Concerning DOTE's credibility on this issue, one House aide quipped: "We have fire and storm emanating from memos, but when it comes to making a really tough decision, the lion becomes a mouse." (7:46) Taken together, the AMRAAM, DIVAD, and B-1 controversies raise urgent questions about the effectiveness, accuracy, and adequacy of IOT&E.

Are IOT&E assessments effective, accurate, and adequate? Unfortunately, the Air Force is ill-equipped to answer this question. Currently, there is no formal review of IOT&E assessments in light of later operational experience with a weapon--no procedure for checking IOT&E predictions against reality. Instead, weapons testing is challenged and defended in an emotionally charged atmosphere with little objective data; a situation not conducive to unbiased evaluation. Emotionalism and polemics are not good ways to judge testing, and some kind of objective evaluation is important to ensure IOT&E is doing the job.

An objective evaluation of IOT&E could produce several benefits. Conceivably, it would highlight testing problem areas and guide changes in organization or technique to solve them. Furthermore, implementation of an objective evaluation system would show critics that OT&E management is effective and concerned with improvement. Evaluation also has the potential to improve the credibility of IOT&E assessments. At the very least, an evaluation of the operational testing conducted for each new weapon system would provide a feedback step currently missing in the acquisition process--a step obviously required for any hope of future IOT&E improvement. After all, it's difficult to improve if current IOT&E performance is unknown. The unknown accuracy and adequacy of IOT&E contributes to the weapons acquisition controversies mentioned earlier in the chapter.

The chapter began with a common concern: Will the new weapons work? The question is clearly related to weapons testing and the fact that no way exists to objectively measure testing effectiveness. Proposing a way to evaluate IOT&E effectiveness is the purpose of this report. The rest of the chapter elaborated on the need for this evaluation method. It contrasted the official purpose of testing with the real-world problems of the B-1, DIVAD, and AMRAAM. The debate over the performance of these weapons is reason enough to examine the effectiveness of operational testing. Finally, some possible benefits of IOT&E evaluation were listed.

Chapter Two will briefly review some of the history of operational testing, a record fraught with reorganization and turbulent change. This restless search for effective operational testing further supports the need for an objective evaluation method.

CHAPTER TWO

HISTORICAL OT&E: A RESTLESS SEARCH

> We regard the creation of the testing and evaluation group as of the utmost importance, since we believe most of our previous failures to be prepared for wars. . . would have been thoroughly exposed had an adequate program of testing and evaluation existed. (13:26)
>
> - President's Scientific Advisory Committee, 1970

> Keep on going and chances are you will stumble on something. (2:177)
>
> - Charles F. Kettering

The history of operational test and evaluation is a turbulent chronicle filled with disputes over various issues. Judging by the number of changes, the issue causing the most disagreement was how to organize for operational testing. Who was to do it, and who should supervise it? Tracing the organizational development of operational testing leads through a bewildering maze of command and staff structures. This chapter concentrates on the pattern of organizational change in OT&E history. The pattern is significant since a high frequency of change is expected when a poorly operating system lacks appropriate feedback. In the case of OT&E, managers knew they had to change something; they just didn't know what. Inappropriate changes led to unforeseen problems eventually requiring still more changes. Although difficult to prove, the lack of appropriate feedback may be partially responsible for 30 years of organizational flux. This unfortunate pattern began when operational testing started to split away from traditional testing in the 1930s.

One of the first organizations expressing an interest in separate operational or tactical testing was the Air Corps Tactical School (ACTS). In the 1930s, the school was one of the lead agencies developing the emerging air power doctrine proposed by Douhet, Trenchard, and Mitchell. (20:45) Interested in how new airplanes could be used to tactically execute the doctrine, the school naturally wanted to begin testing. However, the ACTS desire to test sparked an immediate controversy with the traditional test agency at Wright Field. (13:10) In 1934, a study group appointed by the Secretary of War, the Baker Board, recommended that an independent test unit be set up at the ACTS. (13:9) No action was taken and the controversy continued until 1939, when the Air Corps created a dedicated test unit, the 23rd Composite Group, under the Air Corps Board. (13:10) With this action, the Air Corps separated operational testing from developmental testing done at Wright Field--the first shot of an organizational war lasting 30 years.

Operational testing was off and running on its own, but not without growing pains. In 1940, to make room for pilot training, the Air Corps transferred the 23rd to Orlando. (13:11) "Moving the 23d (sic) to Orlando created an unsatisfactory situation--the 23d still did the majority of its testing at Eglin Field, but remote from its headquarters at Orlando and from the Air Corps Board at Maxwell." (13:11) In 1941, the 23rd Composite Group left the Air Corps Board and became part of the new Air Corps Proving Ground at Eglin Field, a group charged with tactical testing. (13:12-13) Complexity grew as new organizations were added in 1942. In that year, as part of a massive reorganization of the new Army Air Forces, the Pentagon's Directorate of Military Requirements was created to facilitate the incorporation of "combat lessons" in the new aircraft. (13:13) But the reorganizers also saw a need for still another testing group. They created the Army Air Force School of Applied Tactics at Orlando, Florida, to teach combat-proven tactics to new aviators and "test the tactical suitability" of aircraft already tested at the Proving Ground. (13:14) The Orlando school was the third agency charged with some type of testing, and the second performing operational testing.

For a nation at war, three independent testing agencies proved overly cumbersome. Finally, in a 1943 consolidation, both the School of Applied Tactics and the Proving Ground were reassigned to the Army Air Force Board, reporting to the Directorate of Operations, Commitments, and Requirements in Washington, DC. (13:15) However, problems continued until 1945, when it seemed "the system continued to work only because of the cooperation of the various commanders involved." (13:17) In 1946, responsibility for all operational suitability and tactical employment testing was transferred to the Army Air Force Proving Ground Command. (13:18-19) But somehow, the Air University inherited the test oversight responsibilities of the defunct AAF Board. "Besides their academic training and research responsibilities, Air University was responsible to `plan and supervise the development and testing of new and improved methods and techniques of aerial warfare; and to approve, activate, and designate test agencies and monitor all projects involving tactical unit testing.'" (13:19) Unfortunately, the Air University had no association with the Proving Ground Command or its resources.

When General Fairchild began to gather the resources needed to fulfill Air University's testing role, General Quesada, Commander of the Tactical Air Command, violently objected. (13:19) He believed that operational testing belonged with the commands, not with the academics of Air University. General Spaatz agreed and barred Air University from the testing business. (13:19) In 1947, as the Army Air Forces became the US Air Force, developmental testing belonged to the Materiel Command, operational suitability and tactical development testing belonged to the Air Proving Ground Command, and the operational commands performed operational effectiveness testing. Unfortunately, problems continued since "the Air Proving Ground Command, operating in conjunction with, but separate from, the Air Materiel Command and the operational commands, could not satisfy all observers in its role, nor could it represent the operational commands properly. Rapid technological advancement and increasing costs provoked misgivings about how research and development was conducted." (13:20) Even as a separate service, the Air Force was unable to end spasmodic organizational change in the operational testing business.

Between 1947 and 1970, there were several major changes in the organization of OT&E. In 1957, the Air Proving Ground Command was shorn of its independence and absorbed into the Research and Development Command. (13:23) Problems proliferated, and in 1964, a special Air Staff office was created to monitor OT&E. (13:25) But by this time, the operational commands responsible for effectiveness testing were otherwise occupied with the growing war in Southeast Asia.

Vietnam stressed the inefficient, confusing operational testing system to the breaking point. To fight the war, the Air Force needed new systems on the ramp as soon as possible and operational testing took time. Total Package Procurement became a popular acquisition technique and committed the Air Force to production of new weapons without sufficient OT&E. "Costs soared, systems suffered long delays, and many systems experienced reliability and maintenance problems after deployment."(13:25) In 1970, the President's Scientific Advisory Board gave the Air Force failing marks for acquisition: "It became clear that system failures, high acquisition costs, and extensive post-production system modifications could be attributed to inadequate OT&E and, in some cases, to the complete lack of OT&E prior to production."(13:26) As a solution, the Blue Ribbon Defense Panel recommended the creation of a testing office at the Secretary of Defense level. (13:26) Reorganization was still the preferred fix for testing problems.

From 1939 to 1970, major reorganizations scrambled operational testing units every few years. However, the result was not a highly efficient, involved operational testing organization responsive to field requirements. Instead, after 30 years of alternative wiring diagrams and command structures, a presidential panel had pronounced the acquisition system a failure due to inadequate OT&E. Apparently the changes, although frequent, didn't work. Today's managers should be concerned about this historical pattern of change for a couple of reasons.

First, many of the changes in operational testing were made after problems showed up in wartime. New weapons either weren't incorporating the lessons learned in combat, or testing was taking too long and not providing the effective, suitable aircraft needed to do the job. Significantly, today's systems are untested in combat. Is IOT&E doing a good job, or is it unintentionally masking deadly deficiencies? A war would provide answers, but leaves a lot to be desired as a feedback tool.

Secondly, history shows there is no easy fix for complex OT&E problems. Obviously, when a system must be changed again and again, the changes aren't working. Judging from the great number of changes made, improving OT&E is no trivial task. For one thing, large changes in any complex structure are likely to lead to unforeseen consequences. This is particularly true if the manager has difficulty pinning down the exact cause of the problem. The fact that eight different investigative boards worked on weapons testing in the 1970s is testimony to the difficulty of the problem. (11:2) One could pessimistically conclude from history that changes in OT&E organization will continue forever, each new change resulting in undesirable outcomes.

In summary, the first 35 years of OT&E history are characterized by recurring organizational change. Struggling to improve the value of operational testing, managers tried various organizational schemes. In some years, OT&E was subordinate to developmental testing; at other times, OT&E was done by operational commands outside the acquisition system. It became obvious in 1970 that all the changing had not improved Air Force OT&E. In fact, the Vietnam war exposed several examples of complete OT&E failure. All the years of changing had led only to more problems--problems aggravated by combat.

Today, testers can't be dependent on combat to evaluate IOT&E. Complex testing issues demand high-resolution feedback that shows the exact nature of each problem. Only by fixing the specific problems, can testers avoid changes that bring unforeseen consequences. It's high time a method was developed that could provide such feedback before the next war starts. Taking the first step toward that feedback technique, Chapter Three defines OT&E's present-day mission and the challenges to that mission.

CHAPTER THREE

OT&E MISSION AND CHALLENGES

> . . . From 1980 through 1984, DoD itself slowed down or stayed the production of twenty-six weapon systems upon discovering deficiencies during operational testing. (5:41)
>
> - AIR FORCE Magazine

Previous chapters discussed the need for objective feedback on IOT&E effectiveness. This chapter is for readers unfamiliar with operational testing. Naturally, measuring the effectiveness of any process requires complete familiarity with the process and its goals. Accordingly, this chapter lays the foundation of the Operational Testing Effectiveness Evaluation Method (OTEEM) by reviewing the OT&E mission. The prospective evaluator must also know what challenges OT&E is likely to face along the way. The OTEEM should measure IOT&E mission accomplishment with particular attention to the possible deficiencies caused by these challenges. The first step is to review the mission of OT&E.

The mission statement from Chapter One applies to both developmental and operational testing: "Their primary purposes are: to identify, assess, and reduce the acquisition risks; to evaluate operational effectiveness and operational suitability; to identify any deficiencies in the system; and to ensure that only operationally effective and suitable, supportable systems are delivered to the operating forces."(17:2) Air Force Regulation 80-14 defines some of these terms.

> **Acquisition Risk.** The chance that some element of an acquisition program produces an unintended result with adverse effect on system effectiveness, suitability, cost, or availability for deployment.
>
> **Operational Effectiveness.** The overall degree of mission accomplishment of a system used by representative personnel in the context of the organization, doctrine, tactics, threats (including countermeasures and nuclear threats), and environment in the planned operational employment of the system.
>
> **Operational Suitability.** The degree to which a system can be satisfactorily placed in field use, with consideration being given to availability, compatibility, transportability, interoperability, reliability, wartime usage rates, maintainability, safety, human factors, manpower supportability, logistic supportability, and training requirements.

Maintainability. A measure of the time or maintenance resource needed to keep an item operating or restore it to operational. . . status. Maintainability may be expressed as the time to do maintenance. . . as a usage rate of manpower resources. . . as the total required manpower. . . or as the time to restore a system to operational status. . . .

Reliability. The probability that an item will perform a required function under specified conditions for a specified period of time or at a given point in time. Also expressed as the average time an item will perform a specified function without failure.

Critical Issue. Those aspects of a system's capability, either operational, technical, or other, that must be answered before a system's overall worth can be estimated, and that are of primary importance to the decision authority in deciding to allow the system to advance into the next acquisition phase. (17:34-37)

With these definitions in mind, the specific function of OT&E is: "to ensure that only operationally effective and suitable systems are delivered to the operating forces."(17:7) It does this by "identifying, assessing, and reducing" the possibility that something unexpected will have a negative effect on some characteristic of the system. Contrast OT&E's concern for the operating forces with the purpose of Development Test and Evaluation (DT&E): "That testing and evaluation used to measure progress, verify accomplishment of developmental objectives, and to determine if theories, techniques, and materiel are practicable; and if systems or items under development are technically sound, reliable, safe, and satisfy specifications."(17:34) DT&E emphasizes feasibility and specification compliance, while OT&E is concerned with predicting, verifying, and improving the capabilities and characteristics of an operational weapon. OT&E has the following specific objectives:

a. Evaluate the operational effectiveness and operational suitability of the system.
b. Answer unresolved critical operational issues.
c. Identify and report operational deficiencies.
d. Recommend and evaluate changes in system configuration.
e. Provide information for developing and refining:
    (1) Logistics and software support requirements for the system.
    (2) Training, tactics, techniques, and doctrine throughout the life of the system.
f. Provide information to refine operation and support (O&S) cost estimates and identify system characteristics or deficiencies that can significantly affect O&S costs.
g. Determine if the technical publications and support equipment are adequate.
h. Assess the survivability of the system in the operational environment. (17:7)

There are three types of operational testing used to achieve the above objectives: Qualification Operational Test and Evaluation (QOT&E), Initial Opera-

tional Test and Evaluation (IOT&E), and Follow-on Operational Test and Evaluation (FOT&E).

All three types have important functions in the operational test mission. QOT&E is primarily concerned with modifications to existing equipment, or introduction of off-the-shelf equipment that requires no special research and development. (17:3) This report concentrates on IOT&E and FOT&E. These two kinds of operational testing are actually different test phases conducted on the same weapon system. IOT&E is performed prior to production for major new systems requiring research and development. FOT&E is further operational testing performed after the production decision is made--in fact, throughout the lifetime of the fielded weapon system. (17:3) While IOT&E and FOT&E examine many of the same characteristics and share some objectives, they have different purposes.

IOT&E's purpose is reflected in the OT&E mission statement. Making sure that only effective and suitable weapons get to the ramp is a two-step process. First, the operational tester must be able to distinguish the weapons that aren't effective and suitable; and second, report before a production decision is made. The use of operational testing to support decision milestones was introduced in the 1970s. (11:9) Today, the primary purpose of IOT&E is to provide information for decision makers on operational effectiveness and suitability at each decision milestone in the acquisition process. (17:3) Operational testers are increasingly involved in earlier stages of development, providing data on the operational value of proposed weapons, as well as an operational perspective in the development process. (12:9)

Like IOT&E, FOT&E's primary purpose is determined by its timing in the acquisition process. In a classic acquisition program, FOT&E starts after the production decision is made. Therefore, its goal is no longer oriented toward decision making. Instead, FOT&E strives to improve the weapons system or the way it's used. In the words of AFR 80-14: "It is used to refine estimates made during IOT&E, to evaluate changes made to correct deficiencies found in prior T&E, and to identify additional deficiencies." (17:4) Also it helps ". . . to find out whether the system can meet changing operational requirements; to develop or refine employment tactics; to determine the system's operational effectiveness and suitability characteristics. . . and to refine doctrine and training programs." (17:4) FOT&E refines pre-production IOT&E estimates so users can more efficiently employ the weapon. FOT&E is necessary because of several challenges that cause uncertainty in IOT&E results.

The operational tester faces numerous challenges. These include excessive emphasis on cost and schedule, lack of realism in testing, politics or a lack of independence, and the changing threat. These challenges may cause IOT&E estimates to fall wide of the mark. A brief discussion of each challenge and how it might affect a weapon system should prove useful in designing a measurement system to judge test effectiveness. Basically, OTEEM will measure how much the aggregate of these challenges affects a particular test program. The first of the challenges, excessive emphasis on cost and schedule, can cause a number of problems.

Operational testing of nonproduction-representative equipment or lack of sufficient operational testing are indicative of excessive emphasis on cost and schedule. This undesirable situation is sometimes unavoidable if the immediate need for the system is overwhelming. As Mr. Jack Krings, DOTE, says: "In some cases, the operational effectiveness may be secondary. Sometimes you have to buy a scarecrow--it won't kill many birds, but it'll keep a lot of them away."(4:52) He went on to say: "It's vital to get something out as a deterrent, and maybe you can fix it after it's out there. . . . That doesn't sound like very good policy in terms of being very firm about operational requirements, but sometimes it's just a more practical way."(4:53) When there's a schedule crunch, IOT&E test directors may be asked to test hand-built FSD hardware rather than wait for production-representative systems. Unfortunately, test teams that use such shortcuts may misjudge critical characteristics like reliability and maintainability. If unpleasant surprises in any of the system characteristics are traced to production line changes, then it's possible production-representative systems were never tested. A different but closely related consequence of cost-and-schedule mania is insufficient build-up testing.

	IOT&E	FOT&E
MAIN PURPOSE	Decision making	Improve system/Use of system
SECONDARY PURPOSE	Improve system/ Estimate use data	Refine estimates of IOT&E
PRIMARY OBJECTIVES	Learn effectiveness and suitability Answer critical issues Identify/report deficiencies Assess survivability	Recommend/evaluate changes Identify/report deficiencies Refine operating info for logistics, tactics, training, etc.
SECONDARY OBJECTIVES	Obtain operating info Assist tech order/support equipment development Recommend/eval. changes	Tech ord[illegible]/support equipment eval Refine estimates of effectiveness, suitability, survivability

Table 1. OT&E Purposes and Objectives

Insufficient operational or developmental testing can have disastrous results. As a weapon system approaches the end of IOT&E or begins FOT&E,

testers often run comprehensive operational tests as a kind of "final exam." Since many new and uncontrolled factors like inexperienced crews or maintenance are often present during FOT&E or late IOT&E, it's more difficult to trace the exact cause of a malfunction. Moreover, insufficient build-up operational or developmental testing can cause painful questions about system reliability after an unexpected system failure. Late in FSD, decision makers expect the system to be fairly refined, and news of the failure, coupled with the uncertainty of its cause, can lead to further program delays and cutbacks. Therefore, major problems suddenly occurring in the later stages of IOT&E or early FOT&E may be related to insufficient testing caused by too much emphasis on cost and schedule.

Like insufficient testing, lack of realism may also lead to alarming revelations when the chips are down. Some realism will always be lacking in operational testing. For example, it's inappropriate to fire live surface-to-air missiles at a B-1B just to test its countermeasures. Until the system is used in an operational environment, undetected problems may lurk in the design. Unfortunately, the test ranges and techniques used in IOT&E may be used again to test the system in FOT&E, never revealing these hidden problems. If system failures show up after initial use in the field, suspect a lack of realism in IOT&E. The next challenge, politics or the lack of independence, is popular with the press.

At least one researcher sees the history of OT&E as a search for independence. (13:--) The three systems briefly discussed in Chapter One are examples of alleged lack of independence. If a test program really did suffer from this malady, test reports might not include much negative information. Statements like "insufficient data exists but simulations of projected system capabilities indicate" signal problems with independence. However, since equipment problems don't have politics, hidden malfunctions will inevitably show up when the system reaches the field. The final challenge, the changing threat, exists because the acquisition process takes time.

CHALLENGE	EFFECT
Cost and Schedule	
- nonproduction equipment	-production related defects
- insufficient testing	-unsuspected major failure in late IOT&E or early FOT&E
Lack of Realism	-failure in initial field use
Politics/Lack of independence	-numerous unpredicted major failures
Changing Threat	-obsolescence when reaching field

Table 2. OT&E Challenges and How They Affect Weapon Performance

New weapons are designed to counter the projected threat. However, as full-scale development continues over a period of years, the threat may change. To some degree, DIVAD was a victim of this process. (12:14) Unfortunately, testers and/or the contractors may not be aware of the new threat developments. Obsolescence also results when a system takes longer in development than anticipated. Newly fielded systems checkmated by enemy threat development are casualties of this challenge. This short list of problems, summarized in Table 2, is by no means a complete list of the challenges facing operational testers, but gives some idea of how difficult OT&E can be. With all the potential challenges out there, it makes sense to find out how much they really affect testing.

This chapter discussed the tasks that today's OT&E must accomplish. IOT&E and FOT&E have similar objectives, but IOT&E's emphasis is on information for decision makers. FOT&E's emphasis is on improving a weapon or its employment. The different objectives were divided into primary and secondary categories in Table 1. Several challenges were discussed, including over-emphasis on cost and schedule, lack of realism and independence, and the changing threat. The purpose/objectives summary in Table 1 and a knowledge of the different challenges summarized in Table 2 provide the basis for the evaluation method developed for IOT&E in Chapter Four.

CHAPTER FOUR

THE OPERATIONAL TESTING EFFECTIVENESS EVALUATION METHOD: OTEEM AND IOT&E

It is error only, and not truth, that shrinks from inquiry. (18:15)
- Thomas Paine

Previous chapters addressed the testing controversy, the related need for objective feedback, the historical flux in testing organization, and IOT&E's current role in weapons system acquisition. This chapter introduces OTEEM, a method for obtaining IOT&E feedback. First, the desirable features of OTEEM are discussed. Next, OTEEM methodology is explained and used to examine the IOT&E program of an actual weapons system. This example is illustrative only. The detailed evaluation of a test program using a refined OTEEM is beyond the scope of this report.

DESIRABLE FEATURES

There are four features or characteristics that OTEEM should possess. The first of these is goal orientation. The IOT&E primary goals covered in the last chapter were: (1) learn effectiveness and suitability, (2) answer critical issues, (3) identify and report deficiencies, and (4) assess survivability. Since the IOT&E secondary goals are more participatory in nature and have less impact on acquisition decisions, OTEEM concentrates exclusively on the IOT&E primary goals.

OTEEM must also be sensitive to the damaging effects of the OT&E challenges. Recall that the four main challenges were: (1) overemphasis on cost and schedule, (2) lack of realism in testing, (3) politics/lack of independence, and (4) the changing threat. Since the effects of these challenges are usually apparent when a weapon becomes operational, OTEEM should consider information gathered from the field. However, such information gathering must be practical and cost effective, the next characteristic.

The data necessary to support OTEEM must be readily available and inexpensive to obtain. Overburdened operating commands won't spend a lot of effort on a project that doesn't yield immediate operational benefits. Moreover, the method must be inexpensive in light of increasing budget cuts. Therefore, OTEEM should use only information that's already available and cheap to assemble.

Using this cheap, readily available data, the OTEEM should provide a universally applicable summary of test program effectiveness. OTEEM should be capable of evaluating the entire spectrum of acquisition programs, from gas masks to strategic bombers. In this way, OTEEM will facilitate test program comparison and reveal broad trends and relationships. The "big picture" made possible by comparison of OTEEM results should help managers determine the overall health of OT&E. Universal applicability is the last of the four desirable OTEEM features, including goal orientation, cheap and available data, and challenge sensitivity.

METHODOLOGY

To satisfy the above requirements, OTEEM compares "snapshots" taken at different times in the weapons system life-cycle. The IOT&E final report, a summary of IOT&E predictions and assessments, provides the first of these snapshots. The second snapshot is the field experience with the production weapon system summarized in the FOT&E Phase One final report. Since FOT&E and IOT&E already examine many of the same parameters, comparison of the final reports should be easy. Table 3 is a summary of how OTEEM's report comparison method meets the desired characteristics.

CHARACTERISTIC	OT&E FINAL REPORT COMPARISON
Goal Orientation:	Final reports emphasize the four primary objectives of IOT&E. OTEEM will compare goal-related dimensions.
Challenge Sensitivity:	The problems show up in FOT&E. OTEEM uses FOT&E data.
Inexpensive Available Data:	FOT&E already gathers the exact data required. Both OT&E reports address same areas.
Universal Applicability:	Four primary IOT&E objectives general enough to apply to almost any system.

Table 3. How Report Comparison Satisfies Desired Characteristics

OTEEM evaluates IOT&E by comparing IOT&E and FOT&E assessments in five dimensions related to the IOT&E primary goals: effectiveness, suitability, critical issues, deficiency reporting, and survivability. OTEEM uses specific procedures for each of these dimensions.

The OTEEM effectiveness dimension measures the accuracy of the IOT&E weapons system effectiveness assessment. Weapons system effectiveness is a composite measure of mission accomplishment. The discrete elements contributing to mission accomplishment are different for each weapons system. For

example, the effectiveness elements for a cruise missile may be accuracy, range, terrain-following (TF) capability, and time of arrival (TOA). A fighter aircraft might have slightly different elements: weapon delivery accuracy, combat radius, sustained-"g" turn capability, and speed. IOT&E and FOT&E final reports assess each effectiveness element. In the ideal test program, IOT&E assessments should agree with results determined in FOT&E. The next step is to quantify the agreement or disagreement. For the purpose of this report, the adjective ratings for each element are compared. An element that was satisfactory in IOT&E, but deficient in FOT&E is scored as a disagreement. To arrive at the accuracy rating, simply compare adjectives for each element. For example, if ten effectiveness elements are evaluated, and the reports disagree on two, an 80 percent OTEEM accuracy rating is achieved. The accuracy rating, therefore, expresses the overall correctness of the IOT&E effectiveness assessment based on FOT&E results. Sometimes though, IOT&E fails to assess an element due to insufficient testing, mixed results, etc.

An additional rating, OTEEM completion, expresses the percentage of elements where no IOT&E prediction is made. For example, if out of fifteen elements, five were not rated and five disagreed, OTEEM completion would equal the 10 rated elements divided by the 15 possible elements or 67 percent. OTEEM accuracy would then equal the 5 agreements divided by the 10 rated elements, or 50 percent. The completion rating really expresses the degree to which weapons system effectiveness is known after IOT&E. In this case, the status of only 67 percent of the elements was known at the end of IOT&E. Together, OTEEM accuracy and completion make up the IOT&E effectiveness assessment. The same approach is useful in the next dimension.

Suitability has well-defined elements common to many different test programs. Recall from Chapter Three that these elements include availability, compatibility, transportability, interoperability, reliability, maintainability, safety, human factors, and logistics supportability. Many of these factors can be further broken down to smaller components. For example, logistics supportability includes, manpower, technical data, training, and wartime usage rates. (17:35) Many suitability elements can be quantified with measures like Mean Time Between Critical Failure (MTBCF) or Mean Time To Repair (MTTR). Suitability is scored in OTEEM accuracy and OTEEM completion using the same techniques as OTEEM effectiveness. The next dimension uses a similar approach.

IOT&E and FOT&E final reports list critical issues. OTEEM approaches the critical issue dimension two ways. First, what percentage of the issues were answered in IOT&E; and second, were the answers right? Unfortunately, after listing the issues, the final reports may never explicitly answer them. Instead, issue answers are often implied in the report text or summary. Therefore, critical issue accuracy can only be judged by inference. The OTEEM critical issue dimension includes percentages answered and accurate. So far, effectiveness, suitability, and critical issue dimensions have all shared a common accuracy/completion approach. The next dimension, deficiency reporting, requires a different comparison technique.

The deficiency reporting dimension can be seen as an expression of weapon system maturity at the end of IOT&E. The more mature a weapon is when

produced, the fewer critical deficiencies show up in FOT&E. OT&E final reports divide deficiencies into two categories, mission critical and nonmission critical (other). Different labels are used in different programs, but critical deficiencies impact basic effectiveness/suitability, while other deficiencies have a lower level of urgency. OT&E final reports document the number of each type of deficiency. As a weapon system matures, the number of new deficiencies should decrease. Guided by this assumption, OTEEM expresses FOT&E deficiencies as a percentage of IOT&E deficiencies. OTEEM has a Critical Deficiency Reduction Percentage (CDRP) and an Other Deficiency Reduction Percentage (ODRP). For example, if IOT&E reports 32 critical deficiencies, and FOT&E reports 21, CDRP = 66 percent. Unfortunately, until a data base is gathered from many test programs, it will be hard to say whether a particular percentage is good or bad. The final dimension is more difficult to quantify.

Survivability could be expressed several ways. Some of the possible alternatives include system performance against specific threats, or probability of penetration when opposed by a range of different threats. Survivability estimation techniques and results are often highly classified and not available for analysis. Because of this, exact techniques/examples are beyond the scope of this report, but comparison of the various elements (specific threat systems, or aggregate profiles) between IOT&E and FOT&E should yield survivability dimension accuracy and completion percentages. The survivability dimension is the last component of the OTEEM assessment. Next, a sample application helps illustrate the OTEEM in action.

APPLICATION

Any example must be general enough to avoid specific classified element values. Again, the purpose of the example is to illustrate the technique and stimulate thought, not to judge the effectiveness of a particular program. A fair evaluation using OTEEM would require much more in-depth analysis and probably a classified format. The Air-Launched Cruise Missile was chosen for the OTEEM application exercise.

The Air-Launched Cruise Missile (ALCM), officially designated the AGM-86B, is a strategic weapon system procured in the late 1970s and early 1980s. IOT&E on the ALCM extended from 23 April 1979 to 31 March 1980. (14:1) It was conducted in conjunction with a fly-off between two contractors and consisted of 10 launches and 10 captive carries for each contractor. (14:1) Captive-carry missions simulate missile flight with the missile connected to the aircraft pylon. The winning contractor, Boeing Aerospace Company, was awarded the contract, and FOT&E was conducted between April 1980 and May 1981. (15:ii) During FOT&E, eleven launches were conducted with an unspecified number of captive carries. (15:ii) It's important to note that the ALCM is not a perfect example of IOT&E supporting milestone decisions. A critical need for the system forced a production decision before IOT&E was complete. (5:45) Fourteen areas were examined in both IOT&E and FOT&E. These areas were separated into the OTEEM dimensions below. Tables 4-7 list the raw data extracted from the reports for later calculation of the OTEEM ratings. Survivability was not included due to classification. In these tables, "S" = satisfactory, "U" = undetermined, "D" = deficient. The adjective ratings were based on test

team assessments in the final reports. In a few cases, the adjective rating was evident, but not clearly stated as "satisfactory" etc.

ELEMENT	IOT&E RATING	FOT&E RATING
Reliability	D	D
Compatibility	U	U
B-52 systems	U	U
B-52 range/handling	S	S
Interoperability	U	U
Mission Planning	U	U
Data transfer	U	U
Throughput	U	U
Output Accuracy	U	U
Ease of use	U	U
Availability	S	D
Logistics reliability	S	D
Maintainability	S	S
Logistics supportability	S	D
R&M interface	U	D
Maintenance concept (base/depot)	S/U	S/S
Support Equipment	S	S
Planned supply support	U	U
Transportation, packaging, and handling	S	S
Technical data	D	D
Facilities	S	S
Manpower	S	S
Training	S	S
Maintenance training	S	S
Training suit.	S	S
Human Factors	S	S
Software suitability	U	U
Software maintainability	U/S	U/D
Software useability	U	U
OVERALL SUITABILITY	NO RATING(U)	D

Table 4. ALCM Suitability Dimension

ELEMENT	IOT&E RATING	FOT&E RATING
Accuracy: en route/terminal	S/U	S/U
Range	S	S
Terrain Following (TF)	U	D
Launch envelope	U	U
Time of Arrival (TOA)	S	U
Alternate mission capability	S	S
Operational Test Launch (OTL) payload	D	D
Arm and fuze warhead	S	S
Captive carry missile status	S	S
OVERALL EFFECTIVENESS	U	D

Table 5. ALCM Effectiveness Dimension

ISSUE	ANSWERED IN IOT&E	CORRECT IN FOT&E
a. AGM-86B v. AGM-109, which is most cost-effective answer to AF need?	YES	N/A (assume YES)
b. Tech. performance/design parameters demo'd within appropriate threshold value?	NOT ANSWERED	--
c. Compatible with SRAM and gravity weapons?	NOT ANSWERED	--
d. Does Mission Completion Success Probability (MCSP) match SAC requirement?	NOT ANSWERED	--
e. Can digital terrain data and operational navigation requirements be integrated in effective mission profiles?	YES	YES

Table 6. ALCM Critical Issue Dimension

	IOT&E	FOT&E
CRITICAL DEFICIENCIES:	22	80
OTHER DEFICIENCIES:	89	330

Table 7. ALCM Deficiency Reporting Dimension

The final step is to process the raw data using the techniques explained above. When scoring accuracy, only IOT&E elements listed "S" or "D" count (it's impossible to measure the accuracy of "U"). IOT&E "S" or "D" elements that decay to "U" in FOT&E are scored as disagreements. IOT&E "D" elements that improve to "S" in FOT&E are not considered disagreements--system improvement is the desired consequence of IOT&E deficient ratings. Using the rules and techniques above to reach scoring percentages for each of the dimensions yields the following OTEEM results for ALCM IOT&E. The implications of the ALCM OTEEM results are addressed in Chapter Five.

EFFECTIVENESS:	OTEEM Accuracy = 86%
	OTEEM Completion = 65%
SUITABILITY:	OTEEM Accuracy = 79%
	OTEEM Completion = 53%
CRITICAL ISSUES:	Percent Answered =40%
	Percent Correct =100%
DEFICIENCIES:	CDRP = 364%
	ODRP = 371%
SURVIVABILITY:	Not included due to classification

This chapter began with a discussion of desirable OTEEM characteristics including goal orientation, sensitivity to challenges, accessible and inexpensive data, and universal applicability. The OTEEM report comparison method has all the desirable features. Next, the specific methodology for OTEEM was introduced and applied to the ALCM. Chapter Five discusses various findings highlighted by the OTEEM application and some miscellaneous observations and concerns.

CHAPTER FIVE

FINDINGS

> If truth is beauty, how come no one has her hair done at the library?(18:231)
>
> - Lily Tomlin

Chapter Four developed OTEEM and applied it to the ALCM IOT&E assessment. The application exercise was a trial run designed to uncover OTEEM problems and suggest refinements, the subject of this chapter. The application exercise raises several important issues.

A glance at the ALCM effectiveness and suitability data (Tables 4 and 5) reveals the first problem: the large number of IOT&E undetermined or "U" elements. A possible explanation lies in the unique circumstances surrounding the ALCM program. Test managers planned a limited IOT&E program to evaluate unproven technology in the face of a critical need for the system. When technical problems cropped up in testing, decision makers bought the system anyway, accepting a degree of uncertainty in effectiveness and suitability. Justified or not, small completion percentages like these have an effect on the evaluation. Obviously, it's tough to evaluate assessment accuracy without assessments. In this example, however, OTEEM still demonstrated its worth. OTEEM completion percentages highlighted the large proportion of ALCM IOT&E unknowns, a crucial insight for managers reviewing the program. The FOT&E disposition of these IOT&E unknowns is another important issue.

Eighty-nine percent of the ALCM elements rated for the first time in FOT&E were deficient. What could explain a large percentage of IOT&E unknowns turning up deficient in later testing? One possibility is that test managers, realizing the impact of negative OT&E assessments in today's acquisition system, want an air-tight case before reporting deficiencies. If a degree of uncertainty exists, some test managers may feel the "U" is safer than a qualified "D" in the final report. Unfortunately, such a practice can hide vital information from the decision maker. To monitor this potential problem area, an OTEEM measure showing the FOT&E disposition of undetermined IOT&E elements would be a valuable addition to the method. OTEEM leads to further insights when the disagreements between IOT&E and FOT&E ratings are analyzed.

In the ALCM program, the Time of Arrival (TOA) function was rated satisfactory in IOT&E, but undetermined in FOT&E. Briefly, the TOA function is a guidance computer routine commanding the missile to speed up or slow down in order to make a particular timing profile. Missile performance and aerodynamics are absolute limits on the TOA capability. The IOT&E and FOT&E test teams evidently disagreed over the meaning of the evaluation objective:

"Evaluate the operational capability of the TOA function in the missile computer."(14:30) The IOT&E team wanted to see if the routine worked at the anticipated slow-down or speed-up rate, but the final report for IOT&E admits that the testing of this function was limited.(14:31) The test team did not induce artificial errors, evaluating only naturally occurring timing errors. These errors were "small and did not tax the TOA function's capability."(14:31) TOA testing was also limited by conflicting higher priority DT&E objectives.(23:--) Nevertheless, since the demonstrated TOA speed-up and slow-down was close to the expected value, the IOT&E team rated TOA satisfactory. The FOT&E team used a different philosophy.

During FOT&E Phase One, TOA was rated "U" because the team felt that although the function was correct, TOA's exact capability was unknown. They recommended validation and analysis of the mission planning factors constraining TOA performance.(15:38) Until testing determined the limits of the capability for TOA, the FOT&E team did not feel justified in giving a satisfactory rating. The two approaches were unquestionably different. IOT&E personnel verified the TOA function, while FOT&E team members tried to determine the TOA capability. Clearly, this disagreement would never have occurred if test objectives had been carefully written with no ambiguities, and then followed to the letter. OTEEM analysis proved valuable by highlighting this disagreement and encouraging closer investigation. IOT&E and FOT&E reports also disagreed on ALCM availability.

The reports rated availability satisfactory in IOT&E, but deficient in FOT&E. Again, the problem lay in the wording and interpretation of the test objective. The objective was concisely written as: "Estimate the availability of the AGM-86B weapon system."(14:53) According to the IOT&E report, this meant "apparent availability."(14:55) Apparent availability is the number of missiles apparently available to support an Emergency War Order generation and does not include missiles that are inoperative, but have not been detected.(15:59) Because of these undetected, inoperative missiles, the FOT&E team favored real availability over apparent availability.

> The primary measure of AV [air vehicle] availability is termed "real" availability. Calculating real availability takes into account missiles not mission capable because of undetected failures, such as in the engine, and missiles down for inspection or for maintenance caused by detected failures. . . Real availability is a statistical measure of the number of mission capable missiles at a random point in time.(15:59)

During IOT&E, real availability was indeed below the appropriate threshold, but in the words of the IOT&E report: "Since the evaluation criteria were based on apparent availability of a mature AGM-86B system, availability of the AGM-86B was satisfactory."(14:55) In the IOT&E report executive summary, no distinction was made between apparent and real availability: "Availability, logistics reliability, maintainability. . . are all satisfactory."(14:v) Clearly, this kind of misunderstanding can have an effect on decision making. Another problem area highlighted by the ALCM application of OTEEM concerns the critical issue dimension.

The ALCM IOT&E test program left several of the critical issues unanswered. This happened despite 1979 DoD direction that critical issues were:

> Those aspects of a system's capability, either operational, technical, or other, that must be questioned before a system's overall worth can be estimated, and that are of primary importance to the decision authority in reaching a decision to allow the system to advance into the next acquisition phase. (16:22)

No reason was given in the report for leaving these questions unanswered. Although beyond the scope of this paper, a rigorous investigation into the reasons for the unanswered issues could result in valuable lessons.

The final ALCM insight from OTEEM analysis concerns deficiency reporting. ALCM experienced an alarming 364 percent increase in critical, mission-threatening deficiencies in the field. Is this increase normal, or was ALCM particularly immature when ordered into production? Right now, it's impossible to say. Only comparison with different programs will allow the test manager to get a feel for what is normal. Even without a basis of comparison, however, it seems reasonable to question the production maturity of this weapon.

In summary, OTEEM analysis highlighted ALCM IOT&E problems in completing element ratings, interpreting test objectives, answering critical issues, and detecting deficiencies. At this point, it's important to remember that today, ALCM is an extremely capable weapon system with a front-line role in deterrence. However, this is not to say that the test program could not have stood some improvement. Although OTEEM successfully uncovered problem areas in the ALCM IOT&E program, OT&E final reports will need some improvements if OTEEM is to work. These suggested improvements and other miscellaneous ideas are presented below.

REPORT IMPROVEMENTS

Deficiency Data Should Specify Date Written

It was difficult to determine whether FOT&E deficiency data included write-ups dating from IOT&E. The report should specify the testing phase or date when each deficiency was discovered.

Use of Thresholds

For many elements, test teams award adjective ratings based on numerical thresholds. Reports should clearly specify the threshold values acceptable in each element. If the thresholds change for FOT&E, the fact should be clearly highlighted. A brief explanation of what's "satisfactory" or "deficient," using the threshold values, will help keep things on a quantitative basis where possible.

Test Methodology

The final report should explain the way each element was evaluated. Reports should contain sufficient detail to explain disagreements like real versus apparent availability. OTEEM can't compare apples with oranges. Some reports already contain this type of information in sufficient detail.

Critical Issues

Final reports should clearly address critical issues If issues are left unanswered, the report should explain why. The information preserved by such a practice would prove invaluable for later analysis. The current situation requires reading between the lines and guesswork.

MISCELLANEOUS CONCERNS

Quantitative versus Qualitative

The approach used in this report was to compare qualitative adjective ratings based on quantitative thresholds. Another approach would be to compare the exact numerical value for each element. The problem would be how to lump 30 miles of range difference with 200 feet of accuracy difference and come up with some usable overall rating for effectiveness accuracy. Perhaps this quantitative analysis could best be included as an appendix to the regular OTEEM ratings.

Effect of Changing Missions and Threats

Weapons systems are sometimes used for unforeseen missions against unplanned threats. An example is the use of the high-altitude B-52 bomber for low-altitude weapons delivery. Test programs should not be expected to anticipate the effect of completely different mission roles and threats after the weapon is fielded. To avoid the impact of innovative mission roles, the performance data used to judge the effectiveness of IOT&E should be collected early in the operational life of the weapon. Use of the FOT&E Phase One report fulfills this requirement. If other data is used, it should be collected no later than Initial Operational Capability (IOC) plus two years--a commonly accepted milestone for weapons system maturity.

System Improvements Masking Poor Predictions

Suppose IOT&E predicted that weapons system performance in a particular area would be satisfactory. In this hypothetical example, subsequent improvements, unforeseen at the time IOT&E was conducted, eventually ensured that the predicted performance level was reached. Without the unforeseen improvements, the system would not have reached the predicted level. In this case, an OTEEM comparison of predicted versus experienced performance would not highlight the poor IOT&E prediction. There really is no easy solution to this dilemma, except to note that weapons system improvements are natural and desirable. If improvements happen to mask a poor prediction, at least the weapons system is doing the job at the predicted performance level. The opposite case, a system

performing at a worse level than predicted, would be highlighted by OTEEM and investigated.

IOT&E/FOT&E Gaming the System

Anytime evaluation is used, someone will game the system in order to look good. Gaming OTEEM would be easy. IOT&E personnel could simply make extremely conservative estimates in the different elements of effectiveness and suitability. Since OTEEM only spots FOT&E elements that don't live up to expectations, the conservative IOT&E would look very good. Such gaming would reduce the utility of the IOT&E assessment and mislead decision makers. Fortunately, OTEEM gaming is unlikely because of the time between the start of IOT&E and the completion of FOT&E. Many of the IOT&E folks would have moved on to other jobs before an OTEEM evaluation could ever be made and would have nothing to gain by gaming the system. Nevertheless, assessment confidence intervals might be used to reduce any tendency to make overly conservative estimates. For example: Operational Range = 1000nm (plus or minus 100nm). OTEEM could be changed to highlight any result outside the error band.

IOC Plus Two Year Assessment

Sometimes FOT&E Phase One is too early to get a good feel for system performance. As mentioned above, IOC plus two years is accepted as a general definition of system maturity. At that time, the operating command should have a wealth of experience with the actual operational characteristics of the system. Assembling the data would require tracking down all the various offices that file information on reliability, system accuracy, etc. This information could provide the most valid basis for OTEEM comparison, and would be a valuable addition to OTEEM. It could even be used to evaluate the effectiveness of FOT&E Phase One.

Use of Actual Milestone III Briefing Materials

Since one of the main purposes of IOT&E is to support the production decision, the actual IOT&E assessment briefing given to the decision makers would be of interest. Furthermore, since the IOT&E final report may include information gathered after the production decision, it may not represent the actual estimates provided in the IOT&E milestone III assessment. To ensure this valuable information isn't lost in the shuffle, milestone III IOT&E assessment briefings could be included as an appendix in the IOT&E final report.

Retrofit Information

A weapon requiring a large number of retrofits to reach effective and suitable performance was probably immature when produced. A measure showing how many retrofits are accomplished between milestone III and IOC plus two years might also be a good addition to OTEEM. This measure is closely related to deficiency reporting.

This chapter analyzed some of the issues highlighted by the OTEEM application in Chapter Four. Two findings result. First, OTEEM is clearly capable of

detecting a variety of IOT&E problems in individual test programs. For example, OTEEM analysis underscored the importance of clear and precise test objectives. Secondly, some final report improvements are necessary to facilitate OTEEM use. At the end of the chapter, general concerns were raised addressing topics ranging from gaming the system to the use of retrofit data. Overall, this chapter demonstrates that there are valuable insights to be gained through the application of OTEEM. Chapter Six summarizes this report and makes recommendations.

CHAPTER SIX

SUMMARY AND RECOMMENDATION

This report began with a problem, introduced and applied a solution, and discussed the result. A short review of each chapter brings the entire report into focus and provides a foundation for recommendation.

In Chapter One, the problem was introduced. Despite supposedly thorough testing, there is much debate over the capabilities of new weapons systems. For this reason, critics argue that weapons testing is inadequate or ineffective. Their argument is difficult to dispute, since the Air Force is operating its OT&E system without an objective feedback method. The lack of a feedback system and the atmosphere of controversy surrounding acquisition decisions make it particularly hard to determine what problems exist in OT&E and decide if changes are worthwhile. Since effective operational testing is clearly vital in acquiring effective and suitable weaponry, an objective evaluation technique, like the one proposed here for IOT&E, is needed to provide this feedback.

Chapter Two showed that the history of OT&E is characterized by frequent organizational change as managers searched for ways to procure effective and suitable weapons. However, in 1970, after 30 years of ineffective changes, the operational testing system was pronounced a failure. The tendency for repetitive change evident in OT&E history is symptomatic of a poorly operating system with inadequate feedback. In the past, wars provided sporadic general feedback for weapon testing efforts, but never allowed managers to determine the exact problems. A pattern of spasmodic, ineffective organizational change was the result, and will be the result unless the Air Force adopts an appropriate operational testing feedback technique. Chapters One and Two argue that OTEEM is needed to break this pattern and efficiently, objectively diagnose OT&E problems.

The next two chapters introduced and applied OTEEM. Chapter Three laid the foundation needed by readers unfamiliar with testing and evaluation. An understanding of the terminology and philosophy behind present day OT&E is crucial to an appreciation for the Operational Test Effectiveness Evaluation Method. In Chapter Four, desired characteristics like universal applicability were discussed. Then, a crude version of OTEEM was introduced and applied to a real-world operational test program with surprising results.

Chapter Five discussed these results. For example, investigation of factors emphasized by OTEEM analysis showed that interpretation of test objectives was a problem in ALCM IOT&E. Additionally, the chapter contained suggestions to make reports more conducive to OTEEM analysis. Finally, the

chapter discussed general concerns like how unanticipated weapons system improvements might mask poor IOT&E predictions.

After these five chapters, the reader should recognize that the absence of an objective feedback technique has contributed to the acquisition debate and has historically handicapped Air Force ability to judge the effectiveness of its testing system. However, using data available today, it is possible to devise an evaluation technique based on OT&E primary goals to provide this missing objective feedback. The benefits offered by OTEEM range from microscopic post-mortems of specific test programs, to macroscopic views of broad operational testing trends. Given the importance of operational testing, some form of objective evaluation is an absolute necessity.

RECOMMENDATION

Logically, OTEEM should be implemented by the service organization already charged with IOT&E management or oversight. In the case of the Air Force, the Air Force Operational Test and Evaluation Center (AFOTEC) is the obvious choice. OSD's DOTE may also be interested in the method.

A special group should be formed at AFOTEC to handle OTEEM affairs. The group should begin a trial program, applying OTEEM to selected systems that have reached IOC plus two years. After this study is completed, a finalized OTEEM technique should be established and implemented. A data base of OTEEM results could be then generated and could include such components as a yearly OTEEM report.

For decades, operational testing managers have searched for the key to OT&E success. As controversy rages over increasingly complex and expensive weapons, managers must ensure OT&E is as effective and accurate as possible. OTEEM may finally be a way to optimize OT&E methods, silence the critics, and ultimately ensure that the weapons reaching the ramp really are effective and suitable.

BIBLIOGRAPHY

Books

1. Rasor, Dina (ed). More Bucks, Less Bang: How the Pentagon Buys Ineffective Weapons. Washington, DC: Fund for Constitutional Government, 1983.

2. The Readers' Digest Treasury of Modern Quotations. New York: Readers' Digest Press, 1975.

Articles and Periodicals

3. Air Force Times. 21 September 1987, p. 18.

4. Biddle, Wayne, "How Much Bang for the Buck?" Discover (September 1986), pp. 50-63.

5. Canan, James W. "Testing from Chips to Chocks." Air Force Magazine (February 1988), pp 40-45.

6. Lerner, Michael A., with John Barry. "Sergeant York Musters Out." Newsweek (9 September 1985), p 23.

7. Morrison, David C. "OT&E Fails to Quiet the Critics." Military Logistics Forum (June 1986), pp. 43-46, 63.

8. "Pentagon Weighs Plan to Expand Testing Schedule of Weapons Systems." Aviation Week and Space Technology (9 March 1987), pp. 264-265.

9. Powell, Stewart, and Melissa Healy. "The B-1 Bomber: A Flying Lemon." U.S. News and World Report (24 November 1986), p. 29.

10. Van Voorst, Bruce, "The Pentagon's `Flying Edsel.'" Time (19 January 1987), p. 21.

Official Documents

11. Adams, Ronald M., Maj, USAF. Test Concurrency and the Carlucci Initiatives: When Is More Too Much? Maxwell AFB, AL, 1984.

12. Everly, Kieth W., Maj, USAF. United States Air Force Policy for Operational Test and Evaluation. Maxwell AFB, AL, 1987.

13. Oertel, Robert E., Maj, USAF. Operational Test and Evaluation: The Quest for Independence. Maxwell AFB, AL, 1985.

14. US Department of the Air Force: Air Force Test and Evaluation Center. "AGM-86B Initial Operational Test and Evaluation Final Report on ALCM Competition(U)." Kirtland AFB, NM., 1980.

15. US Department of the Air Force: Air Force Test and Evaluation Center. "AGM-86B Air Launched Cruise Missile Operational Test and Evaluation Final Report(U)." Kirtland AFB, NM., 1981.

16. US Department of the Air Force: Test and Evaluation. AF Regulation 80-14. Washington, DC: Government Printing Office, 1980.

17. US Department of the Air Force: Test and Evaluation. AF Regulation 80-14. Washington, DC: Government Printing Office, 1986

18. US Department of the Air Force. The Tongue and Quill. AF Pamphlet 13-2. Washington, DC: Government Printing Office, 1986.

Unpublished Materials

19. US Department of the Air Force: Air Command and Staff College. "Thinking About War: A Survey of Military Theory," text 00033 R01 8503. Maxwell AFB, AL.

20. US Department of the Air Force: Extension Course Institute (AU). "History of U.S. Air Power, Course 50, vol 1," text 00050 01 8406. Maxwell AFB, AL.

Other Sources

21. Feighery, Col, USAF. OSD, DOTE. Telephone conversation in October 1987.

22. Lloyd, Dr. AFOTEC/RSH. Telephone conversation on 19 January 1988.

23. Pulcher, Larry J., Maj, USAF. ACSC/EPT. Conversation in January 1988 recalling his personal experiences as an ALCM IOT&E test team member.

END
DATED
FILM
8-88
DTIC